# 日本語の複単語表現辞書：JDMWE

首藤　公昭[†]・田辺　利文[†]

日常の自然言語文には構成性 (compositionality) に基づいて意味を扱う事が難しいイディオムやイディオム的な複数単語からなる表現，また，語の強い結合によって成り立つ決まり文句や決まり文句的表現が数多く使われているが，現在の自然言語処理 (Natural Language Processing: NLP) ではこれらに十分な対応が出来ていない．近年，この種の特異性を持つ表現を複単語表現 (Multi-Word Expression: MWE) と名付け，NLP の立場から英語の MWE 全体を俯瞰・考察した論文 (Sag et al. 2002) が端緒となって，その重要性が広く認識されるようになった．しかし，その後の活発な研究にも拘わらず，包括的で信頼性のある言語資源を構築するには至っていない．筆者らは，現代日本語を対象とした概念語相当 MWE 辞書の構築を古くから進めてきており，本論文ではその初版の概要を報告する．本辞書，JDMWE (Japanese Dictionary of Multi-Word Expressions) は主として人の内省に基づき，以下を目標に編纂されている．

1. 典型的なイディオムや決まり文句に限定せず，いわば準イディオム，準決まり文句的表現の候補も採録すること
2. 特定の構文構造に限定せず，広範囲かつ体系的に収録すること
3. 異表記，派生形を網羅すること
4. 構文構造情報を与え，表現の構文的柔軟性にも対処すること

現在の収録表現数は基本形で約 104,000 件であり，記載した異表記，派生形情報を使えば 750,000 表現程度をカバーする．本辞書は各 MWE に依存（木）構造を与えた一種のツリーバンクと見なすことができる．

キーワード：言語的特異性，非構成性，慣用句，イディオム，決まり文句，コロケーション，支援動詞構文，機能動詞結合，四字熟語，連結詞，オノマトペ，擬音語，擬声語，擬態語，比喩，換喩，格言，諺，故事成句，挨拶，呼び掛け，独言，応答，フィラー，クランベリー表現，ツリーバンク，n-グラムデータ，内部修飾句，フレーズベース機械翻訳

# JDMWE: A Japanese Dictionary of Multi-Word Expressions

Kosho Shudo[†] and Toshifumi Tanabe[†]

Since (Sag et al. 2002) is presented, the NLP society has been aware that one of the most crucial problems in NLP is how to cope with idiosyncratic multiword expressions, which occur in authentic sentences with unexpectedly high frequency. Here,

[†] 福岡大学工学部電子情報工学科, Department of Electronics and Computer Science, Faculty of Engineering, Fukuoka University

the idiosyncrasy of expression is twofold in principle; one is idiomaticity, i.e. non-compositionality of meaning and the other is the strong probabilistic boundness of word combination. Thus, many trials to extract those expressions from corpora by using mostly statistical method have been made in NLP field. However, presumably because of the difficulty with their correct extraction without human insight, no reliable, extensive resource has yet been available. Authors recognized the crucial importance of such irregular expressions in around 1970 and started to develop a machine dictionary which contains Japanese idioms, idiom-like expressions and other multiword expressions which consist of frequently co-occurring words. In this paper, we give an overview of the first version of the dictionary, namely JDMWE (Japanese Dictionary of Multi-Word Expressions). It has about 104,000 head entries and is characterized by;

1. the wide notational, syntactic and semantic variety of contained expressions,
2. the syntactic function and structure given for each entry expression and
3. the possibility of internal modification indicated for each component word of the entry expression.

**Key Words**: *linguistic idiosyncrasy, non-compositionality, lexicalized phrase, institutionalized phrase, idiom, collocation, support verb construction (SVC), light verb construction (LVC), discourse connective, discourse marker, four Kanji character word, onomatopoeic expression, metonym, proverb, saying, call, answer, greeting, filler, soliloquy, cranberry expression, tree-bank, n-gram data, internal modification, phrase-based machine translation*

## 1 はじめに

日常の自然言語文には構成性 (compositionality) に基づいて意味を扱う事が難しいイディオムや相当数のイディオム的な複数単語からなる表現，また，語の強い結合によって成り立つ決まり文句や決まり文句的な表現が数多く使われている．しかし，現在の自然言語処理 (Natural Language Processing: NLP) ではこれらには必ずしも十分な対応が出来ていない[1]．

[1] イディオム「目を回す」,「水に流す」，決まり文句的表現「引くに引けない」,「何とは無しに」を市販の良く知られた日英翻訳ソフト 2 種に翻訳させた結果を以下に示す．結果からいずれもこれらの表現を正しく認識していないことが推定される．

| | |
|---|---|
| 彼はそれを聞いて目を回した | A 社；He turned his eyes hearing it. |
| | B 社；He heard it and turned eyes. |
| 私は過去を水に流す | A 社；I throw the past into water. |
| | B 社；I pass the past in water. |
| 彼は引くに引けない | A 社；He ..pull.. is not closed. |
| | B 社；He cannot pull to pull. |
| 私は何とは無しにそれを見た | A 社；I regarded it as what nothing. |
| | B 社；..it was. was seen very much..me.. |

近年，このような特異性のある複数単語からなる表現を複単語表現 (Multi-Word Expression: MWE) と名付け，英語の機械処理の立場からその全体像を俯瞰し，対応を考察した論文 (Sag et al. 2002) が端緒となって，NLP における MWE 処理の重要性が広く認識されるようになった．これを受け，（国際）計算言語学会 (Association for Computational Linguistics: ACL) は 2003 年以降，MWE に関するワークショップをほぼ毎年開催しており，活発な議論が行われている．しかし，これまでの研究にはなお，以下の様な基本的な問題点が残っている．

1. 複合名詞 (Noun Compound: NC), 動詞･不変化詞構文 (Verb-Particle Construction: VPC), 動詞・名詞構文 (Verb Noun Construction: VNC)，イディオム (Idiom) など，限られた構文，意味の表現だけを対象とする研究が多い．
2. 典型的なイディオム，典型的な決まり文句などを対象とする研究が多く，意味的非構成性や要素語の共起に特異性を持つと認められるそれ以外の表現が顧みられていない．
3. コーパスから MWE を自動抽出する研究において，基準となる表現集合が不備なために再現率を的確に検証することが難しい．

筆者らは，機械翻訳研究 (首藤 1973) の経験からフレーズベースの訳出が必要であること，一般の NLP にも複数単語からなる特異的な表現を総括的に資源化しておくことが不可欠であることを認識し，現代日本語におけるそれらの候補を収録した辞書の構築を目指してきた．本論文ではその初版の概要を報告する．以後，この辞書を JDMWE (Japanese Dictionary of Multi-Word Expressions) と呼ぶ．本辞書は上記の問題を解消し，日本語の特異的複単語表現の基準レキシコンを与えることを目標に，主として人の内省によって編纂されている．編纂においては以下の点に留意した．

(1) NLP に有効と思われる，出来るだけ広範な MWE 候補を体系的に整理・提示すること[2]
具体的には，イディオム（慣用句），決まり文句（常套句），慣用的な比喩表現，機能動詞結合（一部），支援動詞構文（一部），クランベリー表現，四字熟語，格言，諺，擬音・擬声・擬態語表現，複合語（一部），呼びかけ表現，応答表現等を対象とする．以後，これらの表現および外国語でこれらに相当する表現を MWE (Multi-Word Expression) と総称する．

(2) 異表記，派生形をできるだけ網羅すること

(3) 各 MWE に機能情報のほか，構文構造情報を与えることにより，MWE を単語と見なした処理だけではなく，構文的柔軟性（内部修飾可能性）にも対応できるようにすること

現在の収録 MWE は基本形で約 104,000 表現，記載した異表記，派生形情報をすべて適用して見出しを生成すれば 750,000 表現程度をカバーしていることになる．

[2] ただし，固有表現 (named entity)，頭字語 (acronym)，混成語 (blend)，会話調表現，尊敬・丁寧・謙譲表現には現時点では原則として対応していない．他の辞書類やルールによる自動生成等でカバーされることを想定している．

本辞書はMWEごとにスロット付きの依存（木）構造を与えた一種のツリーバンク，あるいは，語の組み合わせに特異性があると同時に纏まった意味・談話上の機能を持つ，構造付きn-グラム ($2 \leqq n \leqq 18$) データセット (syntactically annotated n-gram dataset) と見なすことが出来る．

以下，2. で関連研究を概観し，本研究の位置付けを明らかにする．3. で本辞書に収録した表現について詳しく述べる．4. で辞書形式を簡単に説明し，辞書内容として異表記に関する情報，機能に関する情報，構造に関する情報について順に述べ，例を用いて構造情報と内部修飾句との関係を説明する．5. では既存の大規模日本語n-グラム頻度データとの比較等によって収録表現の統計的性質に基づいた考察を行う．6. で総括と今後の課題を述べてむすびとする．

## 2 関連研究

(Gross 1986) は，フランス語の複合副詞 (compound adverb)，複合動詞 (compound verb) の種類が単独の副詞，動詞の，それぞれ，3.3倍，1.7倍程度存在することを指摘した．また，(Jackendoff 1997) は，英語の日常使用者の持つMWEレキシコンは単語レキシコンと同数規模だと推定されること，(Sag et al. 2002) はWordNet1.7 (Fellbaum 1999) のエントリーの41%がMWEであることを指摘した．日本語でも [動詞＋動詞] 型の複合動詞が動詞の種類の44%を占めることが (Uchiyama et al. 2003) で示されている．この様に日常の自然言語には意外に多種類の複単語表現が使用されており，充実したMWEレキシコンを整備することが重要であることが認識されている．本論文で述べるMWE辞書，JDMWEの基本見出し数は104,000表現であり，(Jackendoff 1997) の指摘した英語におけるMWEの分布も日本語における分布と大差ない事が推定される．

(Sag et al. 2002) は，さらに，英語のMWE全体を俯瞰し，語彙的に纏められる句 (lexicalized phrase) を形態・構文的な柔軟性の度合いによって，固定表現 (fixed expression)，半固定表現 (semi-fixed expression)，構文的に柔軟な表現 (syntactically flexible expression) に分け，慣習的に使われる句 (institutionalized phrase) と合わせて，それぞれのNLPにおける取り扱い方を論じた．具体的には複合名詞 (compound nominal: CN)，固有名詞 (proper name: PN)，動詞・不変化詞構文 (verb-particle construction: VPC)，軽動詞構文 (light verb construction: LVC)，分解可能イディオム (decomposable idiom)，分解不能イディオム (non-decomposable idiom) などの種類ごとに，単語的な扱い (words with spaces approach, holistic approach) と形態，構文，意味上の構成的な扱い (compositional approach) の是非について論じた．(Sag et al. 2002) の指摘の本質は，MWE現象が広範に亘る事，MWEを単語として扱うだけでなく，多様な形態・構文的柔軟性に応じた取り扱いをしなければならないという事であり，その後のMWE研究に多くの示唆を与えた．(Sag et al. 2002) の枠組みによる日本語MWEに関する考察には (Baldwin et al. 2003a) がある．(Villavicencio 2004) は (Sag et al. 2002) の分類に基づき，英語イディオムと動

詞・不変化詞構文を例に，従来の単語辞書をいくつかの表で拡張する形で MWE をデータ化する方法を論じた．本論文の JDMWE も一般の単語辞書や構文解析機との併用を想定しており，内容的に (Villavicencio 2004) の要請の多くを満たしているが，対象とする表現がより広範である点，辞書としての独立性がより強い点，意味と細かな形態・構文的変化に関する情報は未記載であるが，各表現に対して内部修飾 (internal modification) の可能性を記載している点，日本語特有の異表記に対応している点などに違いがある[3]．

NLP 用の MWE 辞書を作成したという報告には限られた形態の表現のみを対象とするものや採録表現数が比較的少ないものが多い．例えば，フランス語の 22 種の構文構造を持つ動詞型 MWE，12,000 個を辞書化した (Gross 1986)，13,000 個の英語のイディオムを構文構造付きでデータベース化した (Kuiper et al. 2003)，ポルトガル語の 10 種の構文構造を持つ動詞型 MWE，3,500 個を辞書化した (Baptista et al. 2004), オランダ語の一般的 MWE，5,000 個に構文構造を与えて辞書化した (Grégoire 2007)，フランス語の 15 種の構文構造を持つ副詞性 MWE，6,800 個を辞書化した (Laporte et al. 2008) などが見られる．そのほか，英語とドイツ語のクランベリー表現をそれぞれ 77 個と 444 個収集した (Trawiński et al. 2008) の報告がある[4]．

日本語 MWE の NLP 向け辞書化に関する研究としては，古くは日本語の機能語性 MWE, 2,500 種を組み込んだ文節構造モデルを提唱した (首藤 他 1979; Shudo et al. 1980) や，約 20,000 個の概念語性 MWE 集を作成した (首藤 1989) がある．また，機能語性 MWE の異表記，派生表記を生成する階層的な手法を考案し，これによって 16,771 表現（341 見出し）の辞書を編纂した (松吉 他 2007) の研究がある．日本語イディオムに関しては，市販の数種の慣用句辞典から 3,600 個の慣用句を収集して NLP の立場から考察を加えた (佐藤 2007) がある．イディオム，準イディオムに対して形態的・構文的変化への制約や格要素等，修飾句への制約がどこまで意味の曖昧さ解消に利用できるかは今後の重要な課題である．この点を考慮して辞書構築を試みる研究に (Hashimoto et al. 2006) があり，今後の成果が注目される．

本論文の日本語概念語性 MWE を対象とする JDMWE は，収録表現の構文・意味機能が 26 種類にのぼり，上記の各辞書化研究に比べてより広範囲の MWE を対象としていること，特に，イディオム，決まり文句以外に準イディオム，準決まり文句と言える表現候補を多数収録していること，取り扱う構文構造の種類が多彩で，例えば動詞型 MWE の場合，80 種以上の依存パタンを持つ表現が収録されていること，異表記に対応していることなど，従来の研究に見られない特徴がある．

MWE 候補をコーパスから自動抽出する研究が近年盛んであり，例えば，日本語，英語のコ

[3] 形態・構文的変化形，例えば，活用，助詞の交替・挿入・脱落，受動態化や語順の入れ替えによる名詞化の可否等の情報記載については (安武 他 1997) で報告した．

[4] 例えば，「cranberry」の「cran」,「おだをあげる」の「おだ」の様な不明語（クランベリー語）を含む表現はクランベリー表現 (cranberry expression) と呼ばれる．

ロケーション検出を統計的手法と一種のコスト評価で試みた (Kita et al. 1994) の研究，中国語複合名詞の抽出を統計的手法で試みた (Pantel et al. 2001)，既存の意味的タグ付けシステムを統計的手法で補強することによって英語の MWE 候補抽出を試みた (Piao et al. 2005)，形態・構文的柔軟性の少なさを統計的に検出して英語の [動詞 + 名詞] 型イディオム候補抽出を試みた (Fazly et al. 2006; Bannard 2007) の研究など数多い．この種の研究では相互情報量 (mutual information, pointwise mutual information)，$\chi^2$ (chi-squared)，対数尤度 (log likelihood)，KL 情報量 (Kullback Leibler divergence) などが相関尺度 (association measure) としてよく用いられるが，自動抽出における相関尺度と MWE との適合性を比較検討した研究に (Pecina 2008; Hoang et al. 2009) がある．MWE とその要素語のコーパス中でのコンテクストの違いを検出して MWE を認定する研究に (Baldwin et al. 2003b) がある．また，最近は対訳コーパスを利用して MWE 候補を抽出する試み，例えば，英語—ポルトガル語で行った (Caseli et al. 2009)，ドイツ語—英語で行った (Zarreß et al. 2009) などが見られる．一定の概念が言語 A では単語で表現され，言語 B では複数単語の列で表現されるということはしばしば起こる．このとき，言語 B の単語列は MWE である可能性がある．この種の現象を対訳コーパスから検出しようというのがこれらの研究の基本的な考えである[5]．

コーパス中の MWE データはスパースな場合が多く，統計的手法による MWE 捕捉では十分な再現率 (recall) の達成が難しい．また，基準となる表現集合も明確でないため，MWE 自動抽出の再現率評価自体が難しいという問題がある．人の利用を目的としたイディオム辞典類は古くから編纂されてきており，日本語に関しても慣用句を対象とした (白石（編）1977)，(宮地裕（編）1982)，(米川 他（編）2005)，故事ことわざ慣用句を対象とした（尾上（監修）1993; 田島 2002)，四字熟語を対象とした (竹田 1990)，擬声語・擬態語慣用句を対象とした (白石（編）1992) 等々，数多くの成果が出版されているが，これらには典型的表現しか収録されていない場合や，表現の機能，内部構造，異表記，変化形，用法に関する体系的な記述が見られない場合が多く，そのままでは NLP における基準集合とはなりにくい．これらの問題点を緩和するのに JDMWE が役立つことが期待される．

NLP における言語資源の評価は，応用システムの性能向上にどれだけ貢献したかで行うのが現実的であるが，MWE を対象としてそこまで行った研究はまだ多くないようである．この種の研究には，日本語 MWE の主に文字面の情報を使って市販日本語ワープロの仮名漢字変換初回正解率を向上させた (Koyama et al. 1998) の研究，日本語の機能語的 MWE を検出して用いれば，より正しい係り受け解析が実現出来ることを示した (注連 他 2007) の研究などがある．その他の日本語 MWE 処理に関する近年の研究には，複合動詞の多義選択法を考察した (Uchiyama et al. 2003)，複合名詞の機械翻訳方式を考察した (Tanaka et al. 2003) などがある．

[5] 本辞書でも英語への訳出を参考にして選定した表現が多数含まれている．

## 3　採録表現

新聞記事，雑誌記事，小説，随筆，事典・辞書類などの広範な文書から，語の共起に何らかの特異性が認められ，構文・意味・談話上の一定の働きを持つ MWE を，主として編者の内省によって収集・整理した[6]．共起の特異性は，基本的なものとして次の 2 種に注目した[7]．

1. 非構成（イディオム）性
2. 要素語間の強い共起性

### 3.1　非構成性 MWE

要素単語の標準的な機能から表現全体の構造・意味を規則で導くことが難しい，即ち形態・構文・意味上の非構成性 (non-compositionality) を持つ表現，あるいは構成性は成立しているが適用すると過生成 (overgeneration) をもたらすと思われる表現を収録した．細かくは次の様な種類がある[8]．

(1) 意味上の非構成性を持つ表現
通常，イディオム（慣用句）と呼ばれている表現で，例えば，「赤の他人」，「耳を貸す」，「手を抜く」，「足が出る」，「首が回らない」，「顔を売る」，「気を取り直して」，「気が利く」等々である．これら，典型的表現以外にも非構成性には次のような種々のレベルが存在する．

(2) 形態・構文上の構成性が不備，あるいは不明瞭な表現
例えば，文頭で連結詞（文脈接続詞，discourse connective）として使われる「とはいえ」，「にもかかわらず」，「といった訳で」など，挨拶表現の「ありがとう」，「こんにちは」など，サ変名詞性の「見える化」，形容動詞性の「いわずもがな」，副詞性の「しょっちゅう」，掛け声「どっこいしょ」など，また，動詞性の「身につまされる」，連体詞性の「名うての」のようなクランベリー表現，その他，構成的な扱いが過生成を招く表現には，連体詞性の「確たる」，「切なる」，「良からぬ」などがある．

(3) 一部の支援動詞構文 (Support Verb Construction: SVC)
例えば，「批判を加える」，「磨きを掛ける」，「計画を立てる」，「旅行に行く」，「顔をする」，「思いをする」，「ウロウロする」，「心待ちにする」など[9]．

---

[6] 概念語的な働きをする表現を対象とし，「によって」，「かもしれない」などの機能語的働きをする MWE は対象外である．

[7] MWE はこれらの特異性の少なくとも一方を持つ 3 種に分けられるが，辞書中にその種別を明記するには至っていない．特異性の程度は連続的に分布しているため，表現の採否の判断が難しい場合がある．本辞書では，規則・処理系の負担を最小限にするいっぽう，レキシコンを出来るだけ充実させることを念頭に再現率を重視する立場をとった．

[8] (1)–(7) は必ずしも互いに排他的な概念ではない．

[9] 「研究-する」のように全体の意味が規則で求められると思われる表現は対象外とする．

(4) 一部の複合語
例えば,「練り歩く」,「打ち拉がれる」,「積み立てる」,「膝小僧」,「袋叩き」など[10].

(5) 四字熟語
「支離滅裂」,「雲散霧消」,「一心不乱」,「乱離骨灰」,「多事多端」,「危機一髪」,「百鬼夜行」など[11].

(6) 慣用的な比喩表現
例えば,「火ダルマになって」,「命の限り」,「死ぬ程」,「黒山の人だかり」,「血の雨が降る」,「眼を皿にして」,「霧の中にある」など.

(7) その他,意味の構成性に問題が有ると思われる表現
通常イディオムとは呼ばれないが,機械処理において構成性に問題が生じる可能性のある準イディオムと呼ぶべき表現も日常の文書には頻繁に出現する.これらの候補も出来るだけ収録した.例えば,「伝票を切る」,「辞書を引く」,「要求を呑む」,「大学を出る」,「頭が良い」,「風邪を引く」,「思いが熱い」,「命の洗濯」,「約束を反古にする」,「元気が良い」,「扇風機を回す」,「車を転がす」,「カメラを回す」,「だからといって」,「足が速い」等々である.

以上の MWE は,纏まった構文・意味・談話における一定の機能を持つ単語列であり,いずれかの要素単語を同意語,類似語あるいは下位概念の語で置き換えたとき,同じ(類似の)意味にならないか,意味をなさなくなるか,あるいは不自然になるという性質を持つ.例えば,「赤の他人」を「真紅の他人」,「耳を貸さない」を「耳を貸与しない」,「手を抜く」を「手を引き抜く」,「一票を投じる」を「一票を投げる」,「要求を呑む」を「要求を飲用する」などと言い換えたとき,少なくとも元の意味は保存されない.表現の採否は基本的にこの性質に準拠している.

## 3.2 単語間共起性の強い MWE

語の共起性の強い表現は,構文・意味解析において係り先を優先的に決定して解析の曖昧さを低減する処理や語の出現を予測する種々の処理に有効である.ここでの表現には以下のものが含まれる[12].

(1) 共起性の特に強い表現
決まり文句的表現で「風前の灯」,「付きっ切り」,「矢継ぎ早」,「禍転じて福となす」,「雲一つ無い」,「時は金なり」,「願ったり叶ったり」,「手をこまぬく」,「程度の差こそ有れ」,「眼にも止まらぬ早技」,「右肩上がりに」,「不倶戴天の敵」,「灯火親しむ候」など.

(2) 格言,諺,故事成句の類

[10] 「食べ-始める」のように全体の意味が規則で得られると思われる表現は対象外とする.
[11] 四字熟語の機能・用法を本辞書では 4. で述べる枠組みで体系化している.
[12] (1)–(5) は必ずしも互いに排他的な概念ではない.

「急がば回れ」,「一寸の虫にも五分の魂」,「ペンは剣より強し」,「柳に風折れ無し」,「一寸の光陰軽んず可からず」,「初心忘る可からず」,「大海は芥を択ばず」,「石の上にも三年」,「人の振り見て我が振り直せ」,「羹に懲りて鱠を吹く」,「蛍雪の功」など[13].

(3) 擬声，擬音，擬態語を伴う表現
擬声，擬音，擬態語は共起する用言に強い制約のある場合が多い．例えば,「ノロノロと歩く」,「ユルユルと動く」,「グラグラ揺れる」,「グッスリ眠る」,「クルクル回る」,「ポッカリと空く」など.

(4) その他，共起性が比較的強いと思われる表現
「肩の荷を下ろす」,「警鐘を鳴らす」,「景気が上向く」,「烙印を押す」,「悪口を言う」,「メリハリの利いた」,「面子の丸潰れ」,「妄想が膨らむ」など.

(5) 概念に固有の固定的言い回し
特定概念を表現するとき強い単語間の排他的共起性を持つ表現で「情報検索」,「文句を言う」,「女流作家」,「疑惑を生む」,「機械翻訳」,「静寂を破る」等々である[14].

(1)–(4) は，纏まった構文・意味・談話上の機能を持つ単語列 $w_1w_2w_3\cdots w_n$ で，いずれかの要素単語 $w_i$ について，条件付後方出現確率 $p_f(w_i|w_1\cdots w_{i-1})$ あるいは条件付前方出現確率 $p_b(w_i|w_{i+1}\cdots w_n)$ が相対的に高いという確率的な特異性 (probabilistic idiosyncrasy) を持つと思われる表現である．例えば，$p_f$(灯 | 風前の)，$p_f$(無し | 柳に風折れ)，$p_f$(三年 | 石の上にも)，$p_f$(押す | 烙印を)，$p_f$(言う | 悪口を)，$p_f$(鳴らす | 警鐘を)，$p_f$(眠る | グッスリ), $p_b$(手 | をこまぬく)，$p_b$(時 | は金なり)，$p_b$(面子 | の丸潰れ)，$p_b$(初心 | 忘る可からず)，$p_b$(景気 | が上向く)，などは比較的大きいと判断した．(5) は特定概念を表現するという条件のもとで高い単語間共起確率を持つもので，例えば，$p_b$(女流 | 作家)，$p_f$(生む | 疑惑を)，$p_f$(破る | 静寂を) は，それぞれ，$p_b$(女性 | 作家)，$p_f$(起こす | 疑惑を)，$p_f$(壊す | 静寂を) などよりかなり大きいと想像できる．

### 3.3 表現の長さ

本辞書に収録した表現のグラム数と収録数の関係を表 1 に示す．基本的には市販の国語辞典類の単語・接辞を単位としたグラム数である．2〜5 グラムの表現が全体の 90%を超える[15].

## 4 記載情報

本辞書の形式を表 2 に示す．現在，約 104,000 行，9 列（A 欄〜I 欄）からなっている．
以下，各表現に与えた情報について説明する．

[13] (1), (2) は (Sag et al. 2002) の分類における固定表現 (fixed expression) に近い.
[14] (Sag et al. 2002) の分類における慣習的に使われる句 (institutionalized phrase) に近い.
[15] 1 グラムデータは，後述する派生情報によって MWE に変化するため，例外的に見出しに加えた表現である．最長の 18 グラム表現には「天は人の上に人を創らず人の下に人を創らず」がある.

**表 1**　表現の長さと採録表現数の関係

| 表現の長さ（グラム数） | 採録表現に占める割合 (%) | 表現の長さ（グラム数） | 採録表現に占める割合 (%) |
|---|---|---|---|
| 1 | 2.44 | 10 | 0.1 |
| 2 | 18.26 | 11 | 0.03 |
| 3 | 41.18 | 12 | 0.03 |
| 4 | 23.39 | 13 | 0.01 |
| 5 | 8.86 | 14 | 0.01 未満 |
| 6 | 3.29 | 15 | 0.01 未満 |
| 7 | 1.58 | 16 | 0.01 未満 |
| 8 | 0.6 | 17 | なし |
| 9 | 0.23 | 18 | 0.01 未満 |

**表 2**　辞書の形式

| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| いまだかって | いまだ-かって | 未だ-(嘗/曾)(っ) て | D | | DD | | | 否定 |
| いまだかつて | いまだ-かつて | 未だ-(嘗/曾)(つ) て | D | | DD | | | 否定 |
| いまだかってない | いまだ-かって-ない | 未だ-(嘗/曾)(っ) て-無い | Ya | aeb | | | | |
| いまだかつてない | いまだ-かつて-ない | 未だ-(嘗/曾)(つ) て-無い | Ya | aeb | | | | |
| いまだしのかん | いまだし-の-かん | 未だし-の-感 | Mk | | AnoM | 〈no〉-〈de〉 | | |
| いまだしのかんあり | いまだし-の-かん-あり | 未だし-の-感-(有/在) り | Yk | vb20 | | $\phi$-〈de〉 | | |
| いまだしのかんがある | いまだし-の-かん-が. ある | 未だし-の-感-が-(有/在) る | Yv | vb2 | | | | |
| いまだしのかんのある | いまだし-の-かん-の. ある | 未だし-の-感-の-(有/在) る | Tv | vb25 | | | | |
| いまだに | いまだ-に | (未/今) だ-に | D | | Dni | | | |
| いまだもって | いまだ-もって | (未/今) だ-以て | Dv | | DVte | | | |

## 4.1　表記に関する情報

### 4.1.1　平仮名見出し（A 欄）

見出しは平仮名の音表記に基づいている．例えば，「良い」は「よい」と「いい」に，「得る」は「える」，「うる」に，「言う」は「いう」，「ゆう」に適宜読み分けて別見出しとする．また，「はんでぃーたいぷ」，「はんでぃたいぷ」など，外来（カタカナ）語の揺れによる異表記も原則として別見出しとする．見出し総数は約 104,000 件である．

### 4.1.2　字種，表記の揺れ情報（B, C 欄）

B 欄のハイフンおよびドットは語境界を示し，C 欄は字種情報と表記の揺れ情報を与える．例えば，C 欄，「組 (み)-合 (わ) せ」などの括弧は送り仮名などの文字の任意性を，「(有/在) る」，「(良/好/善) い」などの括弧と斜線の組み合わせは文字の選択肢を与える．B 欄，C 欄を組み合

わせれば，殆ど全ての異表記を簡単に生成できる[16].

## 4.2 機能に関する情報（D 欄）

D 欄には表現の文法的機能，あるいは意味的，談話的種別をコード化して記載する．これらの種類とその表現の概数，表現例を表 3，表 4 に示す．コードは各表現の構文木におけるルートノードのラベルに相当する．

表 3 の連結詞性表現 (C)，副詞性表現 (D)，連体詞性表現 (T) のコードに付した添え字 v, a, k は表現がそれぞれ動詞，形容詞，形容動詞を含むことを示す．また，サ変以外の動的名詞性表現 (Md) とは，「する」ではなく「をする」が後接して動詞化する表現である．様態名詞性表現 (Mk) とは，名詞の性質と物事の広義の様態を表す性質とを併せ持つ形容動詞的な名詞表現である．これに対し，形容動詞，準形容動詞性表現 (Yk) は，物事の広義の様態を表すが，名詞性が弱く，格助詞の後接等が出来ない表現である．擬声・擬音・擬態語 (Yo) は，主として G 欄で MWE を派生させる目的で便宜上 MWE の見出し表現に加えている．格言，諺，故事成句 (_P) は，その構造によってさらに 13 種に下位分類されているが，煩雑のため，ここでは説明を省く．_Self，_Call，_Grt，_Res の表現には状況によって意味合いが変わるものがあり，これらのクラスは互いに素ではない．例えば，ねぎらいの呼びかけ表現「お疲れ様です」は，近年，単なる軽い挨拶としてもよく用いられる．

## 4.3 構文構造に関する情報

### 4.3.1 構成単語間の境界（B 欄）

B 欄のハイフンおよびドットは語境界を示すが，ドットは，その直後の単語の独立性が比較的強く，内部修飾句（列）を取り得る事を示す．

表現の単位切りは基本的には市販の国語辞典類の単語単位とするが，異表記を簡潔に表現するため，字種が変化する可能性のある所には区切りを入れた[17].

### 4.3.2 述語の支配構造（E 欄）

収録表現のうち用言を用いた述語性表現 (Yv, Ya, Yk) 約 57,800 とこれらが連体，連用化した様態表現 (Tv, Ta, Tk, Dv, Da, Dk) 約 19,700 は表 3 の例に示した様に格要素等からなる依存構

[16] 例えば，B 欄「き-の-いい-やつ」，C 欄「気-の-(良/好/善) い-(奴/ヤツ)」から，次の 24 種の表記が得られる．「きのいいやつ」，「きのいい奴」，「きのいいヤツ」，「きの良いやつ」，…，「気のいい奴」，「気のいいヤツ」，「気の良いやつ」，「気の良い奴」，「気の良いヤツ」，「気の好いやつ」，「気の好い奴」，「気の好いヤツ」，「気の善いやつ」，「気の善い奴」，「気の善いヤツ」．

[17] 接頭・接尾語，助数詞および造語性の強い使われ方をしている一漢字造語成分は単語と見なして切り離した．
また，活用語尾は原則として語幹から切り離さないが，形容動詞の連用形語尾「に」，「と」，連体形語尾「な」は格助詞との機能・用法上の類似性から助詞相当と見なして切り離した．形容動詞語幹に続く「だ」，「たり」，「なり」は助動詞扱いとした．

**表 3**　収録表現の文法的機能と表現例

| コード | 種類 | 表現数 | 表現例 |
|---|---|---|---|
| C,Ca, Cv,Ck | 連結詞性 | 1,000 | 「とはいえ」,「其れと云うのも」,「残念なことに」,「話せば長い事乍ら」,「裏返して言えば」,「言い換えれば」,「云って置くけど」,「それはそうと」,「念の為だけど」,「裏を返せば」,「其れにしても」 |
| D,Da, Dv,Dk | 副詞（連用修飾）性 | 6,000 | 「後にも先にも」,「嬉しい事に」,「恰好つけて」,「面白がって」,「心から」,「先に述べた如く」,「性懲りも無く」,「不思議と」,「地に足を付けて」,「知らず知らずに」,「時代を問わず」,「厄介な事に」 |
| T,Ta, Tv,Tk | 連体詞（連体修飾）性 | 13,700 | 「物分かりの良い」,「心の広い」,「嘴の黄色い」,「骨の折れる」,「あられもない」,「気の良い」,「品格ある」,「詩情溢れる」,「確たる」,「脇の甘い」,「良からぬ」,「次なる」,「利幅の薄い」,「時宜を得た」 |
| M | 名詞性 | 12,000 | 「赤の他人」,「灰汁の強さ」,「赫々たる戦果」,「事の成り行き」,「飛び石連休」,「先の不安」,「抜ける様な空」,「泣き出しそうな空」,「真っ只中」,「生殺与奪の権」,「燈火親しむ候」,「絶海の孤島」 |
| Ms | サ変名詞性 | 700 | 「逃げ隠れ」,「バカ当り」,「鉢合わせ」,「バトンタッチ」,「誹謗中傷」,「付和雷同」,「一目惚れ」,「見える化」,「東奔西走」,「突然変異」,「泣き寝入り」,「行ったり来たり」,「骨惜しみ」,「自画自讃」 |
| Md | サ変以外の動的名詞性 | 4,000 | 「金の工面」,「付け届け」,「真似事」,「見究め」,「道ならぬ恋」,「見積もり依頼」,「身元の特定」,「民間療法」,「目覚ましい発展」,「持って回った物言い」,「約束事」,「度胸試し」,「取っ組み合いの喧嘩」,「涙ぐましい努力」,「入国手続き」,「根も葉も無いウワサ」 |
| Mk | 様態（形容動詞的）名詞性 | 5,400 | 「二番煎じ」,「二枚腰」,「六日の菖蒲」,「無為徒食」,「昔取った杵柄」,「目一パイ」,「貴方任せ」,「天の邪鬼」,「良い仕上り」,「意地悪」,「一時逃れ」,「一番人気」,「焦眉の急」,「所帯持ち」,「垂涎の的」,「破竹の勢い」,「八方美人」,「不眠不休」,「遣りたい放題」 |
| Yv | 動詞性 | 49,000 | 「目に物を見せる」,「目を回す」,「モゴモゴ云う」,「痒い所に手が届く」,「異臭がする」,「異を唱える」,「数に入れる」,「論理で押す」,「不況から脱出する」,「化けの皮が剥げる」,「危ない橋を渡る」,「性も根も尽きる」,「死中に活を求める」,「打てば響く」,「巧く行く」,「尻尾を巻いて逃げる」,「玉の輿に乗る」,「切って落とす」,「額に汗して稼ぐ」,「どっかと腰を据える」,「どっと疲れが出る」,「海が凪ぐ」 |
| Ya | 形容詞性 | 4,600 | 「目に入れても痛くない」,「らちが明かない」,「うだつが上がらない」,「気が気でない」,「歴史が浅い」,「児戯に等しい」,「それで良い」,「理想からほど遠い」,「肩の荷が重い」,「立つ瀬が無い」,「根も葉も無い」,「枚挙に暇が無い」,「九分九厘間違いない」 |
| Yk | 形容動詞，準形容動詞性 | 3,500 | 「支離滅裂」,「借りてきた猫の様」,「目の玉が飛び出る程」,「一所懸命」,「在り来たり」,「掛け値なし」,「目白押し」,「手応え十分」,「霊験があらたか」,「基本に忠実」,「頭の中が真っ白」,「喰うや喰わず」,「口が達者」,「口癖の様」,「経験豊か」,「木目細か」,「基本に忠実」,「定かでない」,「無病息災」,「火が消えた様」,「進歩的」 |
| Yo | 擬声・擬音・擬態語 | 1,300 | 「ホイホイ」,「ヨチヨチ」,「ヘラヘラ」,「モコモコ」,「モクモク」,「ビチョビチョ」,「クスクス」,「クネクネ」,「スラスラ」,「コテコテ」,「ジリジリ」,「スッポリ」,「ゼーゼー」,「ノッソノッソ」,「ピイピイ」,「ボチボチ」 |

**表 4**　収録表現の意味的，談話的種別と表現例

| コード | 種類 | 表現数 | 表現例 |
|---|---|---|---|
| _P | 格言，諺，故事成句 | 2,300 | 「悪事千里を走る」,「窮すれば通ず」,「三人寄れば文殊の知恵」,「人の振り見て我が振り直せ」,「栄枯盛衰は世の習い」,「歳月人を待たず」,「百聞は一見に如かず」,「義を見てせざるは勇無きなり」 |
| _Self | 自問，独り言 | 170 | 「何という事だ」,「然うだよなあ」,「待てよ」,「厭だなぁ」,「何だったかな」,「困ったなあ」,「しまった」,「やったぁ」,「しめたっ」 |
| _Call | 呼びかけ，掛け声 | 130 | 「済みませんが」,「やあやあ」,「あのもし」,「御苦労さま」,「覚えて居れ」,「静かに」,「ざまをみろ」,「おめでとう」,「チョット済みません」,「どっこいしょ」,「居ない居ないバア」,「お疲れ様です」 |
| _Grt | 挨拶 | 150 | 「いらっしゃいませ」,「お疲れ様です」,「宜しくお願い致します」,「失礼します」,「ご機嫌よう」,「おはよう」,「今晩は」,「お休みなさい」,「じゃあ又」,「じゃあね」,「明けましてお目出とうございます」 |
| _Res | 応答 | 350 | 「ご免なさい」,「然うですかねえ」,「バカも休み休み言え」,「宜しいでしょう」,「やっぱりね」,「何だって」,「冗談もいい加減にしろ」,「すぐ其れだ」,「何て事言うんだ」,「そうなんです」,「了解です」 |

**表 5**　構造パタンと表現の例 （Adv：副詞，N：名詞，p：助詞，Y：用言）

| 依存構造パタンの例 | | コード | 表現例 |
|---|---|---|---|
| [[N + p] + Y] 型 | 名詞 +「を」+ 動詞 | va1 | 「異を唱える」 |
| | 名詞 +「が」+ 形容動詞 | ka2 | 「霊験があらたか」 |
| | 名詞 +「に」+ 動詞 | va3 | 「数に入れる」 |
| | 名詞 +「に」+ 形容詞 | aa3 | 「児戯に等しい」 |
| | 名詞 +「で」+ 動詞 | va4 | 「論理で押す」 |
| | 名詞 +「から」+ 形容詞 | aa5 | 「理想からほど遠い」 |
| [[[[N + p] + N] + p] + Y] 型 | 名詞 +「の」+ 名詞 +「が」+ 動詞 | vb2 | 「化けの皮が剥げる」 |
| | 名詞 +「の」+ 名詞 +「が」+ 形容動詞 | kb2 | 「頭の中が真っ白」 |
| | 名詞 +「の」+ 名詞 +「に」+ 動詞 | vb3 | 「玉の輿に乗る」 |
| [[[Y + N] + p] + Y] 型 | 用言連体形 + 名詞 +「を」+ 動詞 | vc1 | 「危ない橋を渡る」 |
| | 用言連体形 + 名詞 +「が」+ 形容詞 | ac2 | 「立つ瀬が無い」 |
| [[N + p] + [[N + p] + Y]] 型 | 名詞 +「も」+ 名詞 +「も」+ 動詞 | vd1 | 「性も根も尽きる」 |
| | 名詞 +「に」+ 名詞 +「が」+ 形容詞 | ad4 | 「枚挙に暇が無い」 |
| [[Y + p] + Y] 型 | 用言連用形 +「て」（「で」）+ 動詞 | ve1 | 「切って落とす」 |
| | 用言仮定形 +「ば」+ 動詞 | ve2 | 「打てば響く」 |
| [Y + Y] 型 | 用言連用形 + 動詞 | ve3 | 「巧く行く」 |
| [[[[N + p] + Y] + p] + Y] 型 | 名詞 +「を」+ 用言連用形 +「て」（「で」）+ 動詞 | ve5 | 「身を以て償う」 |
| | 名詞 +「に」+ 用言連用形 +「て」（「で」）+ 動詞 | ve7 | 「身に沁みて感じる」 |
| [Adv + [[N + p] + Y]] 型 | 副詞 + 名詞 +「を」+ 動詞 | vec | 「どっかと腰を据える」 |
| | 副詞 + 名詞 +「が」+ 動詞 | vee | 「どっと疲れが出る」 |

造を備えている場合が多い．E 欄はこれらの依存（木）構造，約 80 パタンを va1, aa5, ve7 のように記号化して与える．表 5 に Yv, Ya, Yk の場合の依存（木）構造パタンと表現の例を示す．

#### 4.3.3　末尾の構造情報（F 欄）

MWE に用言とその支配構造が含まれる場合，F 欄には一般形で (α-)*β*と正規表現される英字列を記載する．ここで，α は E 欄に補うべき係り要素がある時にこれを表す．β は述部が他を修飾していたり，複合動詞であったり，助動詞，助詞等を含んでいることなどの情報を与える．例えば，「目玉が飛び出る程」では，全体として名詞性様態表現であることが D 欄で Mk と記され，「目玉が飛び出る」の部分の依存（格）構造が E 欄で va2，すなわち [[N が]V] 型と与えられ，さらに，「飛び出る程」の構造が F 欄で [VV]hodo と与えられる．これらの動詞部を一体化すれば，全体の依存（木）構造が [[[目玉が] 飛び出る] 程] が得られる．本辞書は対象とする表現の構造が多岐に亘るため，辞書の作成・管理上の便宜性を考慮して，構造を分割して記載するこのような方式を採った．述部が単一の用言の場合は β*は空とする．

MWE が用言を含まない場合や含んでいても支配構造を有しない場合，F 欄は，自立語の品詞および接辞を表す大文字と機能語性表現をローマ字表記した小文字列とからなる英字列で構造記述を行う．例えば，「酒は百薬の長」には Mha[MnoM] と記す[18]．品詞記号は，M: 名詞，V: 動詞，K: 形容動詞，D: 副詞，T: 連体詞，P: 接辞とする．

#### 4.3.4　構文的柔軟性（内部修飾可能性）

一般に形態・構文的な柔軟性と意味的非構成性とは相反する関係にあるが，この関係を一律に規定することは難しい．例えば，比較的固いイディオムであっても構文的な柔軟性を持つ場合がある．例を挙げれば，イディオム「油を売る」，「気の置けない」は，それぞれ内部修飾句を取って「油を何時も売る」，「気の全く置けない」と使われることがある．従来の NLP では，イディオムに対して内部修飾句を許さない取扱いが数多く見られる[19]．

本辞書では D, E, F 欄に MWE の骨格構造を与え，B 欄のドットでこの直後の単語が内部修飾を受ける可能性がある事を示す．図 1，2，3，4 に例を示す．

図 1 は動詞性イディオム「手が回る」の例である．D, E 欄の情報，Yv, va2 から表現の骨格となる構造が図の太線の如く与えられ，B 欄のドットによって「手」，「回る」がそれぞれ修飾

[18] $[\cdots[[A_1]A_2]\cdots A_n]$ 型以外の場合のみ括弧 [ ] で句表示を行う．

[19] 市販の日英翻訳ソフト 2 種にイディオムを翻訳させた結果を以下に示す．結果 (1), (2), (5), (6) から「油を売る」，「気の置けない」はイディオムとして認識されていることが判るが，「油を何時も売る」，「気の全く置けない奴」に対する訳 (3), (7), (8) ではこれらのイディオム性が捉えられていない．

| | | |
|---|---|---|
| 彼は油を売る | A 社；He loafs. | (1) |
| | B 社；He idles away his time. | (2) |
| 彼は油を何時も売る | A 社；He always sells oil. | (3) |
| | B 社；He always idles away his time. | (4) |
| 気の置けない奴 | A 社；Intimate fellow | (5) |
| | B 社；A fellow easy to get along with | (6) |
| 気の全く置けない奴 | A 社；Fellow who cannot put nature at all | (7) |
| | B 社；The fellow who cannot place mind at all | (8) |

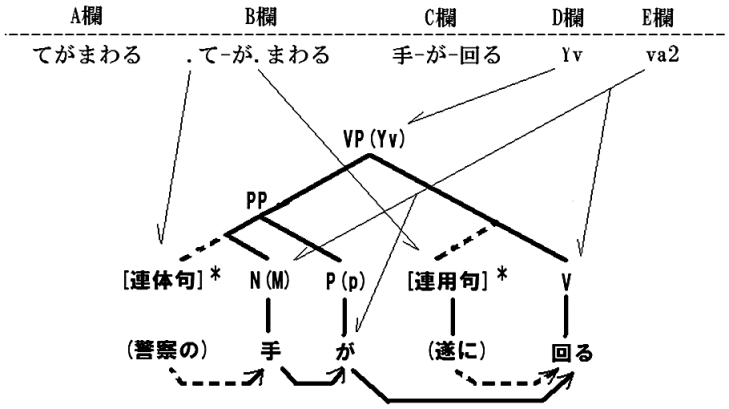


**図 1**　「手が回る」に与えたスロット付き依存（木）構造

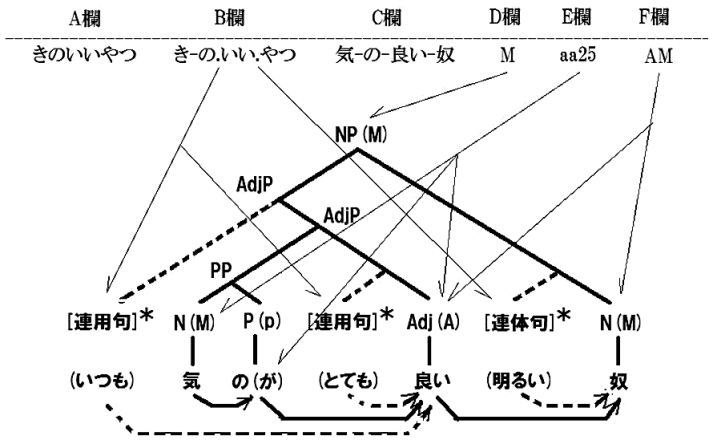


**図 2**　「気の良い奴」に与えたスロット付き依存（木）構造

句（列）を取り得ることが示されている．このことから「警察の手が回る」,「警察の手が遂に回る」のような変化形への柔軟な対応が可能になる．

図 2 は名詞性 MWE「気のいい奴」の例である．D, E, F 欄から太線の骨格が与えられ，B 欄から「いい」,「奴」がそれぞれ修飾句（列）を取り得ることがわかる．これらから，「いつも気のとてもいい明るい奴」などの派生的表現にも対応可能となる[20]．F 欄に記されている AM は表現末尾が形容詞述語による連体修飾構造であることを表し，全体の構造は E の構造を F の構造の形容詞部に埋め込むことで得られる．

図 3 は，副詞（連用修飾）性 MWE,「先に述べた様に」に与えられている構文情報である．ここでも B 欄のドットによって「先に詳しく述べた様に」,「理由を先に詳しく述べた様に」な

[20] E 欄の aa25 は，「が格」支配の形容詞句が「の格」支配の連体修飾型に変化していることを表す．

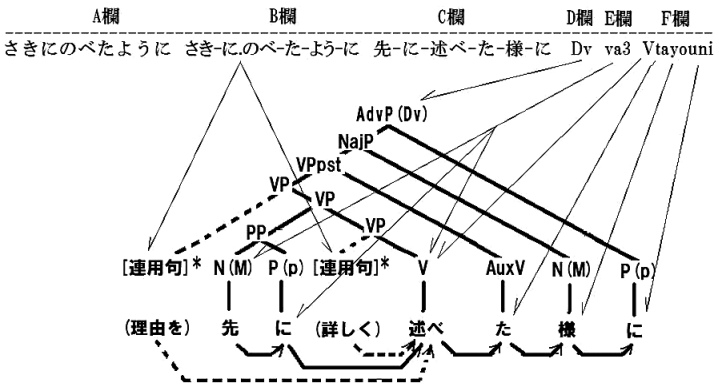


**図 3**　「先に述べた様に」に与えたスロット付き依存（木）構造

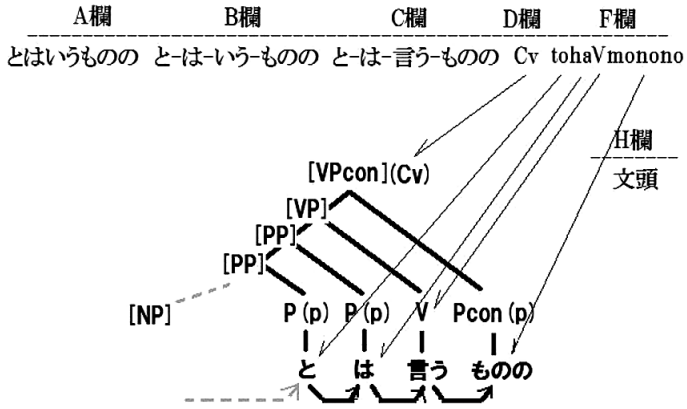


**図 4**　「とは言うものの」に与えた依存（木）構造

どへの対応が可能となる．

図 4 は文頭で用いられる連結詞性の MWE「とは言うものの」の例である．この表現は構成語間の結合の特に強い表現であるため，B 欄にドットが記されていない．図の（不完全な）依存（木）構造は D, F 欄から導くことが出来る．この表現は文頭に位置しなければならないことが H 欄に記されている．

以上の様に，JDMWE は表現ごとに修飾句スロット付き木構造を明記した表現集となっている．

## 4.4　派生情報（G 欄）

形容動詞や形容動詞性名詞，副詞，連体詞など，物事の広義の様態を述べる表現をここでは様態表現と総称する．様態表現は連体，連用，動詞化に関しては用法が様々で，十分な整理を

行っておく必要がある[21]．本辞書では様態表現 (D，Mk，_P，Yk，Yo) に対して G 欄に

〈連体修飾形〉,

〈連体修飾形〉-〈連用修飾形〉 あるいは

〈連体修飾形〉-〈連用修飾形〉-〈動詞形〉

の形式で派生の仕方を記載する．例えば，「我関せず」という表現では，「我関せず-の」，「我関せず-という」，「我関せず-といった」で連体修飾，「我関せず-と」，「我関せず-で」と連用修飾句が派生することを

〈no, toiu, toitta〉-〈to, de〉

と記載する． また，「目玉が飛び出る程」では，「目玉が飛び出る程-の」で連体修飾，「目玉が飛び出る程」あるいは「目玉が飛び出る程-に」で連用修飾，「目玉が飛び出る程-になる」と動詞化することを

〈no〉-〈$\varepsilon$, ni〉-〈ninaru〉

と記す．同様に，擬態語「フラフラ」に対しては，「フラフラ-の」，「フラフラ-した」，「フラフラ-とした」で連体修飾，「フラフラ」，「フラフラ-と」，「フラフラ-して」，「フラフラ-として」で連用修飾，「フラフラ-する」，「フラフラ-とする」と動詞化することを

〈no, sita, tosita〉-〈$\varepsilon$, to, site, tosite〉-〈suru, tosuru〉

と記す．同じ擬態語でも「グングン」では連用句としての「グングン」，「グングン-と」しか有り得ないので，

$\phi$-〈to, $\varepsilon$〉

と表わされる[22]．これら $\phi$-〈to, $\varepsilon$〉 などの派生パタンは約 300 種である．

この種の派生形を別見出しとすれば，見出し数は約 130,000 件程度に膨らむ．

## 4.5 コンテクスト情報

### 4.5.1 文頭側情報（H 欄）

表現が MWE として存立するための制約条件として文頭側コンテクストを指定する．例えば，「顔をする」は単独では用いられず，「嬉しそうな-顔をする」のような連体修飾句が必要であることを〈連体修飾〉と記す．また，「割れになる」は「元本-割れになる」のように，文頭側に連接した名詞による修飾が必要であることを〈名詞連接〉と記す，等々である．この種の条件は約 30 種定めている．

### 4.5.2 文末側情報（I 欄）

H 欄と同様に文末側コンテクストを指定する．例えば，「如何とも」は文末側に「難い」など

[21] その他の用法は相当品詞単語の用法に準じるものとする．
[22] $\varepsilon$, $\phi$, は，それぞれ，空列，用法なしを意味する．

の困難性を表す表現を要求することを〈困難性〉と記載する．同様に，「どの程度まで」に対しては〈疑問〉と記される，等々である．この種の条件は約 70 種定めている．

## 5 考察

収録表現群の統計的性質の一端を探るため，(工藤 2007) の GoogleN グラムデータ（以降，GNG あるいは GNG データと略記する．）との照合を試みた．これは 200 億文からなる日本語 WEB コーパスにおける単語 1〜7 グラムの出現頻度を求めた大規模データである． 対象とした表現は [名詞 $w_1$ + 格助詞 $w_2$ + 動詞 $w_3$] 型の動詞性表現 (Yv) で，格助詞 $w_2$ を「を」,「が」,「に」に，動詞部 $w_3$ を単独の動詞，[動詞 + 動詞] 型複合動詞，あるいは [サ変名詞 + する] 型動詞のそれぞれ終止形に限定した．これらの見出し数は 29,389 個であり，辞書中の B，C 欄の情報で展開した対象表記数は 82,125 個である．これらの $w_1w_2$ 部分の表記数は 13,806 個で，その内 12,120 個が GNG における 2, 3 グラムデータに一致した[23]．これらの表記を前部分列とする GNG の 3，4，5 グラムデータ中で格助詞の直後に上記の種類の動詞（終止形）が出現するものは 1,194,293 個であった．これらの前部分列 $w_1w_2$ ごとに，各動詞の出現頻度を GNG で求めた結果[24]，本辞書データの動詞が GNG で出現頻度第 1 位である場合が 5,787 件であり，対象とした前部分列表記 $w_1w_2$ の $47.74\% = (5,787/12,120) \times 100$ に対して 3.2 で述べた $p_f(w_3|w_1w_2)$ が最大の動詞部 $w_3$ が選ばれていると推定できた．「ちょっかいを出す」,「熱戦を繰り広げる」,「アクションを起こす」などはこれらに該当する．同様に，第 2 位の場合は 1,699 件で 14.02%，3 位は 877 件で 7.24%，4 位は 482 件で 3.98%，等々であった．20 位までの結果をグラフ化して図 5(a) に示す．収録表現は高い条件付き確率のものほど多いというこの結果は 3.2 で述べた MWE 採録の目標から見て妥当なものと思われる．

図 5(a) を累積の比率に改めたグラフを図 5(b) に示す．これから，例えば，本辞書では，対象とする前部分列 $w_1w_2$ の約 80% に対して頻度 8 位までの動詞 $w_3$ が選ばれ，$w_1w_2$ の約 86%に 20 位までの動詞 $w_3$ が選ばれていることなどが分る．GNG データで高い頻度順位の動詞であるのに本辞書で選ばれていないのは，動詞の出現確率に偏りが少なく，絞り込みが効果的に行えないと判断されたためと思われる．また，図 5(b) を外挿すれば，前部分列の 10%強に対して，後接する動詞が GNG では同環境に現れていないことが推定できる．例えば，本辞書に在る「才知に長ける」,「轢き逃げを働く」は GNG に存在しない．このことは，200 億文規模の WEB コーパスであっても，かなりの表現が捕捉出来ない可能性を示唆しており，Zipf の法則におけるロングテール部に対する表現収集の難しさを示すものと考えられる．

[23] $w_1$ が 2 グラムの場合を含む．
[24] GNG データ上の品詞判定には (浅原 2003) の IPADIC 動詞辞書 (verb.dic) およびサ変名詞辞書 (noun.verbal.dic) を用いた．

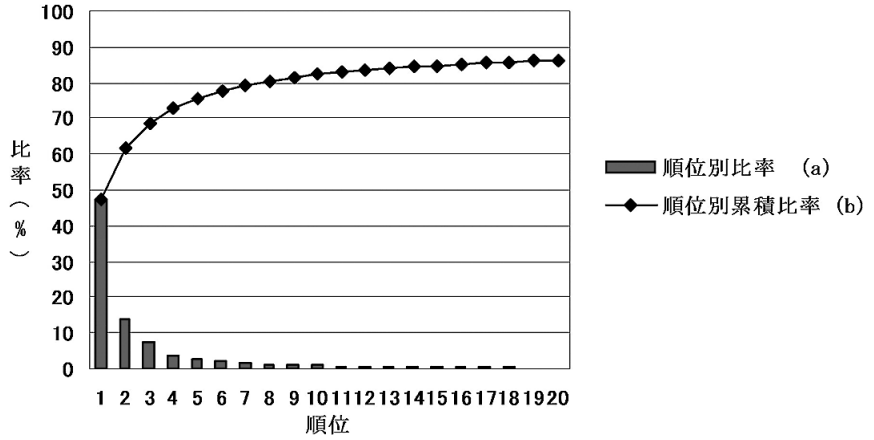


**図 5** [名詞 + 格助詞 + 動詞] 型表現の GoogleN グラムにおける動詞の出現頻度順位別の動詞採録率 (a) と順位別の動詞採録累積比率 (b) （格助詞を「を」,「が」,「に」に限定）

上記 1,194,293 表現の出現頻度の合計は 1,389,568,825 であるのに対し，本辞書データ 82,125 個の出現頻度の合計は 374,718,334 であり，本辞書の表現は GNG の出現数の 26.97%をカバーしている．いっぽう，動詞のバリエーションは GNG で平均 $98.54 = 1,194,293/12,120$ 個であるのに対し，本辞書データでは平均 $5.95 = 82,125/13,806$ 個にすぎない．すなわち，6.04%というコンパクトな動詞の種類で GNG における同環境での出現数の約 27%をカバーしていることが分る．

以上の結果は，限られた形式の表現に対する条件付後方出現確率のみに関するものであるが，採録基準は全体に共通しており，条件付前方出現確率に関しても，また，その他の形式の表現に関しても類似した結果が得られるのではないかと推測している．

本辞書の収録表現が一般の新聞紙上でどの程度使われているかの調査も随時行ってきた．無作為に取った 5 日分の日本経済新聞朝刊第 1 面と最終面における本辞書採録表現の出現比率を表 6 に示す．新聞の 100 文当り本辞書の 73 表現程度が日常的に現れていると推測される．

以上のように，日常の文書ではイディオム性あるいは強い共起性を持つ，比較的少ない種類の MWE が相当高頻度で用いられていることが推測される．

イディオム性データの再現率は，本辞書を利用するシステムの意味構成ルールが明確でない現時点で正しく検証することは難しいが，本辞書データが市販の慣用句辞典類に収録されている表現をほとんど網羅していることは確認済みであり[25]，また，数々の机上実験から弱いイディオム性表現もかなり網羅されていると考えている[26]．

[25] 例えば，(佐藤 2007) が参考にした (宮地（編）1982)，(米川 他（編）2005)，(金田一 他（監修）2005)，(金田一（監修）2005) の慣用句は本辞書に網羅されており，それらの異表記，変化形も相当数収録されている．

[26] ただし，人の内省によってもこの種の表現集合を完全な形で一挙に提示することは難しい．現在，日刊紙の 100 文中に 1 件〜数件程度の新造語その他，登録すべきであろうと判断される表現が出現する．

**表 6** 新聞紙上における 1 文当りの採録表現出現比率 (B/A)

| 日付 | 文数 | | | 採録表現延べ出現数 | | | B/A |
|---|---|---|---|---|---|---|---|
| | 第 1 面 | 最終面 | 計 | 第 1 面 | 最終面 | 計 | |
| | A1 | A2 | A1 + A2 = A | B1 | B2 | B1 + B2 = B | |
| 2009.01.05 | 121 | 142 | 263 | 122 | 63 | 185 | 0.7034 |
| 2009.03.01 | 262 | 133 | 395 | 203 | 110 | 313 | 0.7924 |
| 2009.04.21 | 123 | 141 | 264 | 104 | 115 | 219 | 0.8295 |
| 2009.05.18 | 111 | 137 | 248 | 73 | 78 | 151 | 0.6089 |
| 2009.10.30 | 121 | 168 | 289 | 100 | 110 | 210 | 0.7266 |
| | 計 | | 1,459 | 計 | | 1,073 | 0.7354 |

## 6 おわりに

本辞書は日本語の日常使用者が持っていると思われる言語モデルを「語の慣用」という視点に絞って提示する試みである．表現の選定等は基本的に編者の内省に基づいているため，ある程度の恣意性が入ることは免れない．しかし，5. で見た様に，確率的側面に関しては，大局的には表現の選定に大きな瑕疵は無いものと考えている． 表現の選定に際しては再現率を重視したため，構成性が認められそうな表現や共起の排他性がそれ程高くない表現が採録されている可能性がある．しかし，その様な表現に対しても辞書中に構文構造と内部修飾（分離）可能性を記載しており，それは入力文の通常の構文解析結果を部分的に先取りした情報となっている．この意味で本データは表現レベルの係り受けデータとなっており，これらの表現が機械処理上，障害あるいは無駄になることは少ないと考えている．

本辞書の想定する基本的な応用領域はコンピュータによる日本語の構文・意味・文脈解析であるが，日本語学，日本語語彙・語句論，辞書学，日本語教育等の領域にも参考データを提供できる可能性がある．NLP システムとしては，

1. フレーズベース仮名漢字変換
2. フレーズベース機械（音声）翻訳
3. フレーズベース音声認識
4. 日本語による検索エンジン
5. 日本語による対話システム
6. 日本語読み上げ，仮名振りシステム
7. 日本語教育システム

など，多岐に亘る貢献が期待される．

辞書内容の更なる充実策として以下の点が挙げられる．

a. 並列，反復，対照など，依存以外の構造記載

b. 要素語に対する活用形の記載
c. 形態・構文的変化形，例えば，助詞の交替・挿入・脱落，受動態化や語順の入れ替えによる体言化などへの制約の記載
d. 格要素等，修飾句への構文的，意味的制約の記載
e. 意味上の曖昧さの有無情報の記載
f. 標準的な表現への言い換え情報（含，分解可能性情報）の記載
g. コンテクスト条件として選好 (preference) 条件の記載
h. 「です」,「ます」調，会話調表現の充実
i. 古語，現代語の区別情報の記載
j. 異表記間の優先度情報の記載
k. 条件付き確率，条件付きエントロピー推定値の記載

今後，これらの補強を行って完成度をさらに高めて行く事が望まれる[27].

## 謝　辞

本研究の端緒を与えて下さった故栗原俊彦元九州大学教授，その後，研究上のお世話になった故吉田将元九州芸術工科大学長，長尾真元京都大学総長現国立国会図書館長，ご鞭撻を賜った大野克郎九州大学名誉教授に深甚の謝意を表します． また，有益な助言を頂いた島津明北陸先端科学技術大学院大学教授，翻訳の立場から貴重な意見を頂いた倉骨彰氏，データの収集作業に協力頂いた武内美津乃氏，高丘満佐子氏をはじめとする方々に心から感謝いたします．本論文で検証に用いた IPADIC，Google N グラムデータの関係者の皆様にも深く感謝いたします．

## 参考文献

浅原正幸，松本裕治 (2003). IPADIC version 2.7.0 ユーザーズマニュアル. 奈良先端科学技術大学院大学 情報科学研究科.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003b). “An Empirical Model of Multiword Expression Decomposability.” In *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 89–96.

Baldwin, T. and Bond, F. (2003a). “Multiword Expressions: Some Problems for Japanese NLP.” In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (Japan)*, pp. 379–382.

[27] 本辞書は若干の修正および補強 (a, b) の後に日本語の MWE 解説書と併せてリリース予定である.

Bannard, C. (2007). "A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora." In *Proceedings of A Broader Perspective on Multiword Expressions, Workshop at the ACL 2007 Conference*, pp. 1–8.

Baptista, J., Correia, A., and Fernandes, G. (2004). "Frozen Sentences of Portuguese: Formal Descriptions for NLP." In *Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pp. 72–79.

Caseli, H. M., Villavicencio, A., Machado, A., and Finatto, M. J. (2009). "Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains." In *Proceedings of 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications (ACL-IJCNLP 2009)*, pp. 1–8.

Fazly, A. and Stevenson, S. (2006). "Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations." In *Proceedings of the 11th Conference of the European Chanpter of the ACL*, pp. 337–344.

Fellbaum, C. (ed.) (1999). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Grégoire, N. (2007). "Design and Implementation of a Lexicon of Dutch Multiword Expressions." In *Proceedings of A Broader Perspective on Multiword Expressions, Workshop at the ACL 2007 Conference*, pp. 17–24.

Gross, M. (1986). "Lexicon-Grammar. The Representation of Compound Words." In *Proceedings of the 11th International Conference on Computational Linguistics, COLING86*, pp. 1–6.

Hashimoto, C., Sato, S., and Utsuro, T. (2006). "Detecting Japanese idioms with a linguistically rich dictionary." In *Language Resource and Evaluation, 40-3*, pp. 243–252.

Hoang, H. H., Kim, S. N., and Kan, M. (2009). "A Re-examination of Lexical Association Measures." In *Proceedings of 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications (ACL-IJCNLP 2009)*, pp. 31–39.

Jackendoff, R. (1997). *The Architecture of Language Faculty*. Cambridge, MA: MIT Press.

金田一秀穂（監修） (2005). 小学生のまんが慣用句辞典. 学研.

金田一春彦，金田一秀穂（監修） (2005). 新レインボー小学国語辞典改訂第 3 版. 学研.

Kita, K., Kato, Y., Omoto, T., and Yano, Y. (1994). "A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information VS. Cost Criteria." In *Journal of Natural Language Processing, 1-1*, pp. 21–33.

Koyama, Y., Yasutake, M., Yoshimura, K., and Shudo, K. (1998). "Large Scale Collocation Data and Their Application to Japanese Word Processor Technology." In *Proceedings of the 17th International Conference on Computational Linguistics, COLING98*, pp. 694–698.

工藤拓，賀沢秀人 (2007). Web 日本語 N グラム第 1 版. 言語資源協会.

Kuiper, K., McCan, H., Quinn, H., Aitchison, T., and Van der Veer, K. (2003). "SAID: A Syntactically Annotated Idiom Dataset." *Linguistic Data Consortium 2003T10*.

Laporte, É. and Voyatzi, S. (2008). "An Electronic Dictionary of French Multiword Adverbs." In *Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 31–34.

松吉俊，佐藤理史，宇津呂武仁 (2007). 日本語機能表現辞書の編纂. 自然言語処理, **14** (5), pp. 123–146.

宮地裕（編） (1982). 慣用句の意味と用法. 明治書院.

尾上兼英（監修） (1993). 成語林—故事ことわざ慣用句. 旺文社.

Pantel, P. and Lin, D. (2001). "A Statistical Corpus-Based Term Extractor." In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, Springer-Verlag*, pp. 36–46.

Pecina, P. (2008). "A Machine Learning Approach to Multiword Expression Extraction." In *Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 54–57.

Piao, S., Rayson, P., Archer, D., and McEnery, T. (2005). "Comparing and combining a semantic tagger and a statistical tool for MWE extraction." *Computer Speech and Language*, Elsevier, pp. 378–397.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). "Multiword Expressions: A Pain in the Neck for NLP." In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING2002*, pp. 1–15.

佐藤理史 (2007). 基本慣用句五種対照表の作成. 情報処理学会研究報告, 07-NL-178, pp. 1–6.

首藤公昭 (1973). 専門分野を対象とした日英機械翻訳について. 情報処理, **14** (9), pp. 661–668.

首藤公昭，楢原斗志子，吉田将 (1979). 日本語の機械処理のための文節構造モデル. 電子通信学会論文誌, D62-D-12, pp. 872–879.

Shudo, K. (1980). "Morphological Aspect of Japanese Language Processing." In *Proceedings of the 8th International Conference on Computational Linguistics, COLING80*, pp. 1–8.

首藤公昭 (1989). 日本語における固定的複合表現. 昭和 63 年度文部省科学研究費特定研究 (I)「情報ドクメンテーションのための言語の研究」, 63101005, 報告書.

注連隆夫，土屋雅稔，松吉俊，宇津呂武仁，佐藤理史 (2007). 日本語機能表現の自動検出と統計的係り受け解析への応用. 自然言語処理, **14** (5), pp. 167–197.

白石大二（編） (1977). 国語慣用句大辞典. 東京堂出版.

白石大二（編） (1992). 擬声語擬態語慣用句辞典. 東京堂出版.

田島諸介 (2002). ことわざ故事・成語慣用句辞典. 梧桐書院.

竹田晃 (1990). 四字熟語・成句辞典. 講談社.

Tanaka, T. and Baldwin, T. (2003). "Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing." In *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 17–24.

Trawiński, B., Sailer, M., Soehn, J., Lemnitzer, L., and Richter, F. (2008). "Cranberry Expressions in French and in German." In *Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 35–38.

Uchiyama, K. and Ishizaki, S. (2003). "A Disambiguation Method for Japanese Compound Verbs." In *Proceedings of ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 81–88.

Villavicencio, A. (2004). "Lexical Encoding of MWEs." In *Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pp. 80–87.

安武満佐子，小山泰男，吉村賢治，首藤公昭 (1997). 固定的共起表現とその変化形. 言語処理学会第 3 回年次大会発表論文集, pp. 449–452.

米川明彦，大谷伊都子（編） (2005). 日本語慣用句辞典. 東京堂出版.

Zarreß, S. and Kuhn, J. (2009). "Exploiting Translational Correspondences for Pattern-Independent MWE Identification." In *Proceedings of 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications (ACL-IJCNLP 2009)*, pp. 23–30.

## 略歴

**首藤　公昭**：1965 年九州大学工学部電子工学科卒業．1970 年九州大学大学院工学研究科電子工学専攻博士課程満退．工学博士．同年，福岡大学工学部電子工学科講師．現在，同電子情報工学科教授．機械翻訳，自然言語処理，特に日本語処理に関する諸研究に従事．ACL, 情報処理学会各会員．

**田辺　利文**：1993 年九州大学工学部情報工学科卒業．2000 年九州大学大学院システム情報科学研究科知能システム学専攻博士課程修了．博士（工学）．同年，福岡大学工学部電子情報工学科助手．現在，同学科助教．自然言語処理の研究に従事，人間の気持ちを理解できるシステムをつくることが当面の目標．ACL, 情報処理学会，電子情報通信学会，人工知能学会各会員．

（2010 年 2 月 10 日　受付）
（2010 年 5 月 31 日　採録）